scientific reports

OPEN

Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor

Qiu Cheng[1], Zhan Chunhong[2] & Li Qianglin[1]✉

Small-scale distributed water treatment equipment such as sequencing batch reactor (SBR) is widely used in the field of rural domestic sewage treatment because of its advantages of rapid installation and construction, low operation cost and strong adaptability. However, due to the characteristics of non-linearity and hysteresis in SBR process, it is difficult to construct the simulation model of wastewater treatment. In this study, a methodology was developed using artificial intelligence and automatic control system that can save energy corresponding to reduce carbon emissions. The methodology leverages random forest model to determine a suitable soft sensor for the prediction of COD trends. This study uses pH and temperature sensors as premises for COD sensors. In the proposed method, data were pre-processed into 12 input variables and top 7 variables were selected as the variables of the optimized model. Cycle ended by the artificial intelligence and automatic control system instead of by fixed time control that was an uncontrolled scenario. In 12 test cases, percentage of COD removal is about 91. 075% while 24. 25% time or energy was saved from an average perspective. This proposed soft sensor selection methodology can be applied in field of rural domestic sewage treatment with advantages of time and energy saving. Time-saving results in increasing treatment capacity and energy-saving represents low carbon technology. The proposed methodology provides a framework for investigating ways to reduce costs associated with data collection by replacing costly and unreliable sensors with affordable and reliable alternatives. By adopting this approach, energy conservation can be maintained while meeting emission standards.

Rural domestic sewage is characterized by unstable water quality and quantity, dispersed discharge and low pollutant concentration[1]. To address these challenges, small-scale distributed water treatment equipment has become widely used in the field of rural domestic sewage treatment due to its rapid installation and construction, low operation cost, and strong adaptability[2]. In recent years, the sequencing batch reactor (SBR) process has emerged as a promising option for rural domestic wastewater treatment. When compared with other processes, SBR can effectively withstand organic load impacts, has flexible operation modes, produces good effluent effects, and achieves better nitrogen and phosphorus removal effects[3–6].

However, constructing accurate simulation models for rural domestic wastewater treatment can be challenging due to the non-linearity and hysteresis characteristics exhibited by the SBR process[7,8]. The non-linear problems in sewage treatment refer to the complex, diverse, and non-linear relationships that arise from the interactions of various chemical reactions, biological reactions, and physical effects during sewage treatment.

Artificial intelligence, including machine learning, has been applied to sewage treatment processes to effectively solve non-linear problems. Machine learning encompasses a range of methods, such as neural networks and support vector regression, which can be used to analyze and model the complex data generated during sewage treatment. This has effectively improved sewage treatment efficiency and quality while reducing treatment costs.

[1]Department of Material and Environmental Engineering, Chengdu Technological University, Chengdu, China. [2]Huicai Environmental Technology Co., Ltd., De Yuan Zhen, Pidu District, Chengdu, Sichuan, China. ✉email: polymermacromole@126.com

Artificial neural network (ANN) is a mathematical model that simulates the behavior of animal neural networks, and performs distributed and parallel information processing. ANN has become widely used in predicting sewage discharge, as it can adjust the interconnections among a large number of internal nodes to process complex information within the system[9–13].

In addition to using artificial neural network (ANN) methods, other techniques such as linear regression (LR), support vector regression (SVR), and neuro-fuzzy network methods have also been used in pollutant removal technology to predict changes in pollutant concentrations or other process parameters[14–19]. These methods (as shown in Table 1) have been proven effective in modeling the complex relationships between various factors and predicting pollutant concentrations, which helps to optimize the performance of the treatment process.

However, despite these models[14–17] performed quite well, their processing or environment is idealized. Most of them use simulated experimental conditions. Once in a real engineering case, due to its complexity, the model's performance will not be so outstanding[18,19]. In addition, in these cases[14–19], there are many types of input data, such as DO, pH, conductivity, BOD, COD, TN, etc., which increases the workload or difficulty of data acquisition. For example, there is a significant lag in the measured data of DO sensors; BOD can only be measured using biochemical method and cannot be accurately measured online using sensors due to significant hysteresis. However, although COD measurement can be conducted online, the chemical online measurement method requires strict control of measurement conditions and continuous addition of reagents. The COD sensor method mainly uses optical sensors, which are significantly affected by the chromaticity and turbidity of wastewater. Moreover, the COD sensor is expensive for dispersed small equipment, making it difficult to popularize. Therefore, it is necessary to develop sensors with stable and accurate data collection, long service life and cheap price to replace sensors with poor stability, short service life and expensive prices.

However, the traditional ANN algorithm is based on the asymptotic theory, the empirical risk approaches the actual risk only when the sample size approaches infinity, so the sample size is far from infinity in practical application, it leads to the problems of poor extrapolation ability, slow convergence speed and local extremum[20–23].

Random forest model is one of machine learning, that has become one of research hotspot in the field of artificial intelligence, which has strong adaptive learning ability and nonlinear mapping ability[24,25]. It is suitable for the simulation of wastewater treatment process with the characteristics of large lag, non-linearity and multi-variable[26,27].

The random forest regression (RFR) is a critical application of the random forest (RF) algorithm, which is a statistical learning theory developed by Breiman[28]. The RFR technique involves using Bootstrap resampling to extract multiple samples from the original data and construct decision trees for each Bootstrap sample. These decision trees are then combined to predict the results, with the final prediction being the average of the predictions generated by all the trees[29].

The essence of the RFR algorithm is multi-decision tree model, which makes prediction by combining multiple decision trees. The algorithm has the advantages of high prediction precision, good generalization ability, fast convergence speed and less adjustment parameters, which can effectively avoid "over-fitting" and is suitable for the operation of various data sets. It is robust to the variable extraction of data sets and suitable for ultra-high-dimensional variable vector space. RFR has been widely used in many fields such as medicine, management and agriculture[30–32].

RFR also makes full use of limited samples and construct multiple decision tree models, which increases the diversity of decision tree and improves the accuracy of the final optimization integration model[33,34]. Table 2 shows related applications of random forest regression.

ANN is a kind of machine learning algorithm that is commonly used for predicting the treatment effect of sewage water[45–49]. However, one of the major weaknesses of ANN is overfitting, which can lead to a reduction in the model's generalizability[50–52]. In contrast, the random forest regression (RFR) model is another machine learning algorithm used for predicting sewage water treatment effects. The RFR model has several advantages, including high prediction accuracy, fast processing efficiency, strong generalization ability and is not easily susceptible to overfitting[53,54]. These features make the RFR model an attractive option for predicting sewage water treatment effects.

References	Variables/inputs	Targets/outputs	Model performance	Model
[14]	pH, time, Initial concentration of Cu(II), Nano zero-valent aluminum dose, stirring rate, and temperature	Cu(II) removal efficiency	MSE:ANN < 10^{-5} LR 0.01 SVR 10^{-3}	ANN, LR, SVR
[15]	Temperature, pH, dissolved oxygen (DO), electrical conductivity (EC), NO_3^-, and PO_4^{3-}	Dry cell weight	Determination coefficient (R^2) 0.983	ANN
[16]	Current intensity (I), pH, Fe^{2+} amount, and initial diazinon concentration	Diazinon removal Efficiency	R^2: 0.994	ANN
[17]	Temperature, pH, time, Initial concentration of Cr(VI), and polyamine/folic acid composite dose	Cr(VI) removal efficiency	R^2: 0.919	ANN
[18]	pH, conductivity, BOD, COD, TN influents	BOD, COD, and TN effluents	R^2: BOD 0.764–0.783, COD 0.926, 0.939, TN 0.941–0.957	Neuro-fuzzy networks
[19]	Ten attributes of filament bacteria	SVI	R^2: 0.78, MSE:6	ANN

Table 1. Methods used in pollutant removal technology to predict changes.

References	Variables/inputs	Targets/outputs	Determination coefficient (r^2)
[35]	Temperature, precipitation, and wind	$PM_{2.5}$ of Yangtze River Delta of China from 2015 to 2020	> 0.9
[36]	County-level census data, natural suitability, and socio-economic factors	population distribution of the Tuojiang River Basin from 1911 to 2010	0.84
[37]	Population, agricultural discharge, domestic discharge, sewage collection and treatment way	COD_{Mn} for the Taihu Lake basin in Zhejiang Province, China	0.78
[38]	Runoff data in the same month of the first three years and the runoff data of the first three months	Runoff data of river in Xiaojin County, China	0.85
[39]	Conductivity, turbidity	Nitrate (89%) Total N (85%) Total P (74%) of the lake george drainage basin of U.S	Nash–Sutcliffe efficiency coefficient (NSE) similarly to the coefficient of determination
[40]	Season, outdoor $PM_{2.5}$ concentration, the number of air cleaners deployed, and the density of gers (traditional felt-lined yurts) surrounding the apartments	Indoor $PM_{2.5}$ concentrations	0.815
[41]	Nitrogen application, agricultural and developed land area, and impervious or developed land in the 100-m stream buffer	Loads of total nitrogen	0.76
[42]	Particulate matter 2.5, soil moisture, and relative humidity	Negative air ion in a warm-temperate region of China	0.931
[43]	Real-time color attributes and the environmental conditions of drying process	Moisture ratio of drying date fruit chips	0.976
[44]	Temperature, Wind speed, relative humidity	Ozone concentration in Malaysia	0.970

Table 2. Application of random forest regression (RFR).

Scholars used RFR to predict pollutants concentration in the ambient air[55–59] and urban sewage treatment effect[60–62]. However, there are comparatively fewer studies on the prediction and control of rural domestic sewage treatment effects using the RFR model.

The proposed methodology aims to achieve improved prediction and effective control of the treatment effect of rural domestic sewage through the development and utilization of RFR soft sensor model. By utilizing this approach, it is hoped to establish a reliable and robust soft sensor model that can accurately monitor and analyze key indicators of sewage treatment in rural areas. This will not only facilitate the identification of potential issues and assist in their resolution but also contribute to overall improvements in local ecological conditions and public health standards.

Soft sensor is a commonly used method in process monitoring and control, which estimates the process variable of interest based on the measurements of other variables that are easy to acquire. The establishment of a soft sensor model usually involves selecting relevant input variables, designing the mathematical model, and training the model using historical data. The resulting model can then be used for real-time prediction and control. Soft sensors have been widely applied in various industrial processes such as chemical processes, wastewater treatment and power plants. The advantages of soft sensor include cost-effectiveness, flexibility and ability to handle complex nonlinear systems. Soft sensor has proven to be a valuable tool for process optimization and control[63–66].

Methods

RFR model.

Construction of RFR model. RFR model is an integration algorithm developed on the basis of decision tree theory, which belongs to bagging type[67]. By combining multiple weak learner cart trees and taking the mean value to integrate multiple models, the final result is obtained[68].

The RFR model uses the disturbance of samples and attributes, and increases the "diversity" of the cart tree of the weak learner, so that the final integration result has high accuracy and generalization performance[69]. The RFR model solves practical problems such as small samples, high dimensions and multi-classification, and can handle both discrete data and continuous data[70]. It overcomes the shortcomings of slow convergence speed of neural networks and requires a large number of samples, It also solves the problem of over fitting or under fitting of decision tree, and has good applicability and popularization[71]. Figure 1 shows the diagram of RFR.

Prediction method. The general prediction method of RFR model is:

(1) Randomly take samples from training samples (n × sample) for n times to form a training set (samples were put back after every sampling). Repeat r times to obtain training sets:$D_1, D_2, \ldots, D_r$.

(2) For each training set, k attributes are randomly selected from the attribute set (m × attribute), $k = \log 2m$, and then cart trees are established:$f_1(x), f_2(x), \ldots, f_r(x)$.

(3) The final prediction value of random forest is determined by the average method:$f(x) = \frac{1}{r}\sum_{i=1}^{r} f_i(x)$.

Evaluation index of the model. In order to evaluate the performance of the COD concentration prediction model, mean square error (MSE) and determination coefficient (r^2) are selected as evaluation indexes. The indicators are calculated as follows:

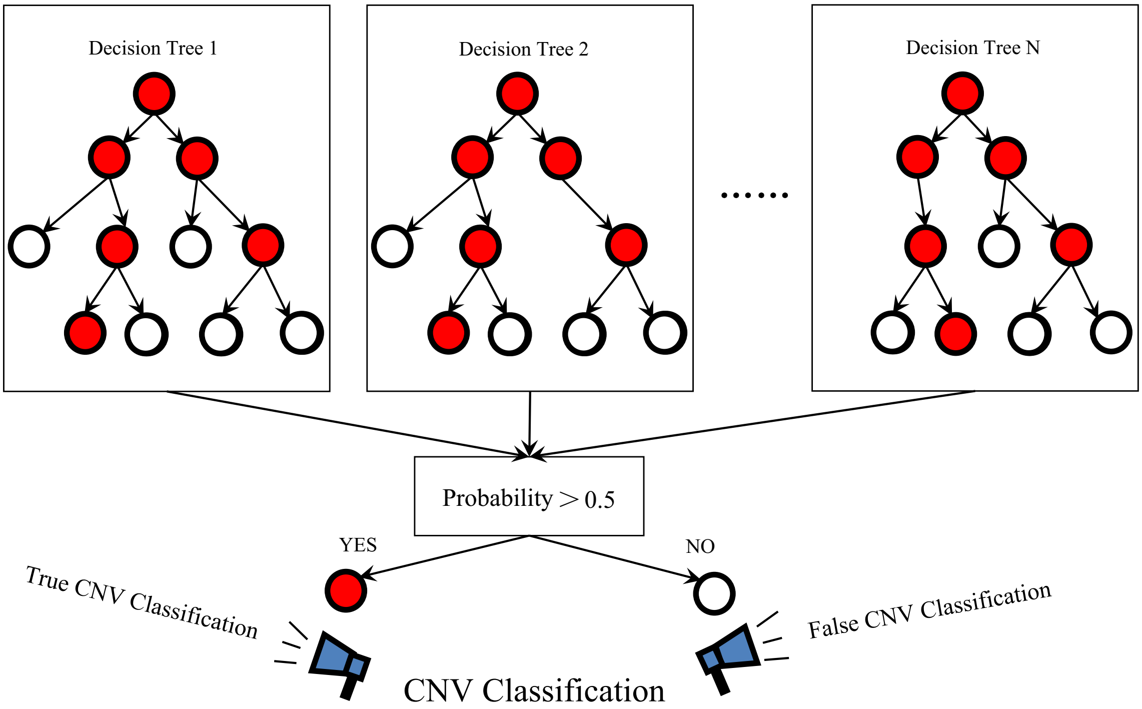

Figure 1. Diagram of RFR model.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

Formula $\hat{y}_i$ for the model predicted value, y_i for the true value.

Characteristic of RFR model. The characteristic of the RFR model were set as Table 3.

Materials and methods.

Structure of SBR. A sequencing batch reactor (SBR), receiving sewage water from a residential area, is prepared in this study. The source of domestic sewage is from the Shuyuan Community in Pidu District, Chengdu, Sichuan, China (longitude: 103.88, latitude: 30.82). The sewage water flowing into the SBR comprised domestic wastewater that had been primary filtered and precipitated. The SBR reactor (stainless steel, 800 mm × 800 mm × 1200 mm) was designed and manufactured. The working volume of the reactor was 0.576 m^3, respectively (Fig. 2). An agitator and an aeration device are installed in the reaction tank.

Sewage water that had been primary filtered and precipitated was pumped into the SBR. This pump was called pump A which made 0.1856m^3 sewage fed to the SBR every cycle. Pump B transported the same volume of water out of the SBR when the cycle ended.

Control. The SBR process is automated and controlled by a PIC (Programmable Integrated Circuit) or a one-chip computer. The cycle, which lasts for 480 min, includes the following stages: 30 min fill in and aeration stage, 330 min oxidation and agitation (alternating aeration and agitation, with aeration lasting for 10 min and agitation 20 min) stage, 60 min settlement stage and 60 min discharge stage. Figure 3 shows time management of the operation of SBR.

Monitoring. Monitoring influent and effluent wastewater samples were taken from the SBR tank and from a collection vessel which allows filtered water go through in order to get rid of the interference of activated sludge.

Filtered pH and temperature were tested by monitoring sensors which manufactured by LuHeng Co. of China (pH:pH sensor LuHeng 6503; temperature:temperature sensor LuHeng 229).

Filtered COD was tested by potassium dichromate method, NH_3-N was tested by Nessler's reagent colorimetry method (SP-756P UV visible photometer of Shanghai spectrum) and TP was tested by spectrophotometric

Characteristic	Value	Function
Number of trees in the forest	100	Refers to the number of decision trees included in the random forest. Increasing the number of decision trees can improve the stability and classification performance of the model, but it will increase computation time. Usually, choosing an appropriate number of decision trees can achieve better results
Number of features to consider when splitting the decision tree each time	$\sqrt{n}$, n = number of input variables	Refers to the number of features considered when each node performs feature selection. Generally, this parameter needs to be set small to reduce the variance of the model. It is usually recommended to set it to the square root of the total number of features, which ensures that different feature subsets are considered when each decision tree splits, increasing the diversity and generalization performance of the model
Criterion for the split nodes of the decision tree	MSE	Specifies the evaluation criteria for splitting decision tree nodes
Maximum depth of the decision tree	10	Controls the maximum depth that the decision tree can grow. A too large depth can lead to overfitting, while a too small depth can result in underfitting. Therefore, this parameter needs to be adjusted appropriately to achieve the best performance
Minimum number of samples required to split an internal node	5	Controls the minimum number of samples required to split each internal node. If the number of samples in an internal node is less than this value, the node will not generate any child nodes, and the branch at this position will stop growing. Setting this parameter value too large may lead to underfitting, while setting it too small may lead to overfitting
Minimum number of samples required to be at a leaf node	3	Controls the minimum number of samples required for each leaf node. For small datasets, this parameter needs to be set smaller to ensure that the model has enough flexibility

Table 3. Characteristic of RFR model in the proposed methodology.

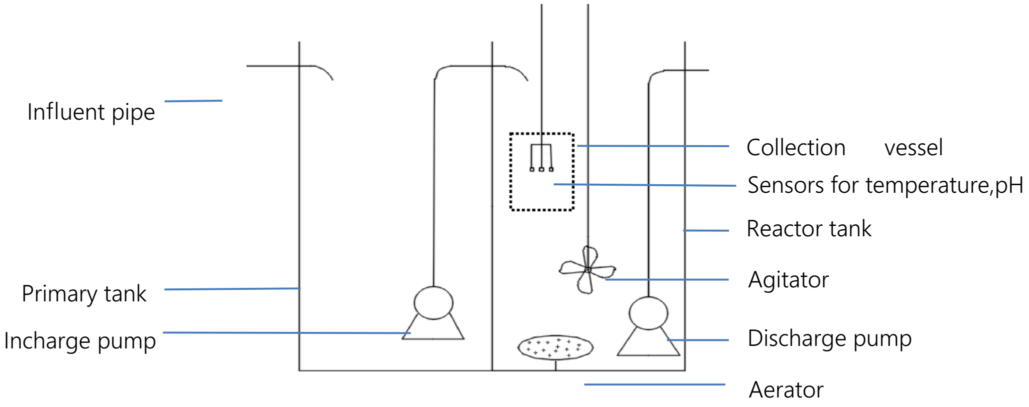

Figure 2. Structure of SBR.

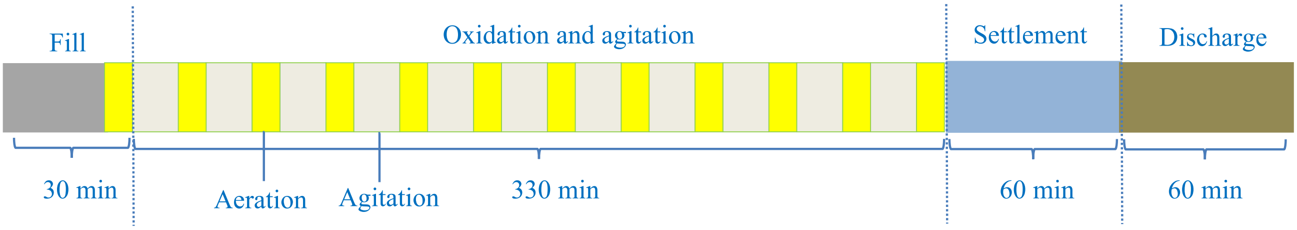

Figure 3. Treatment process of SBR.

detection method (SP-756P UV visible photometer of Shanghai spectrum).pH and temperature were measured at 10 min intervals by sensors.

Sensors were fitted approximately 200 mm below the lowest liquid level within the reaction tank and above any potential sludge blanket that might be formed during settlement. All instruments were calibrated, maintained and operated in accordance with manufacturer' instructions.

Overview of COD, pH and temperature profiles. A typical profile for COD saw an increase in concentrations as influent was mixed with the treated sewage water remaining in the reactor from the previous cycle. COD concentrations peaked soon after the fill phase. Following this peak, COD concentrations decreased due to organic carbon oxidation and nitrification[72]. At approximately 250 min, the rate of decrease in COD concentrations has no more obvious change and continued thus for the rest of the cycle (Fig. 4).

A cyclical rise and fall in pH (Fig. 5) profile during the aeration phase occurred, as the aerator switched on and off, resulting in a peak and trough in each aeration period in pH profile. The increase in pH, corresponding to the aeration-on period, was likely, in this case, to be due to CO_2 stripping[73]. The decreases in pH profile during the 20 min stirring period were likely due to a microbial activity which release carbon dioxide[74].

As found, during the early stage of the cycle, pH fell harder than the later, it is probably because the COD concentration differs: higher COD concentration results in more activity of microorganisms. In general, pH decreases as alkalinity is consumed during the nitrification progress. Denitrification progress causes the overall increase of pH at the medium and end stage probably[75].

$$NH_4^+ + 2O_2 = NO_2^- + H_2O + 2H^+ \text{(pH decreases period)}$$

$$NO_3^- + 5H = 1/2N_2 + 2H_2O + OH^- \text{(pH increases period)}$$

$$NO_2^- + 3H = 1/2N_2 + H_2O + OH^- \text{(pH increases period)}$$

A cyclical rise and fall in pH profiles during the aeration phase occurred, as the aerator switched on and off, resulting in a peak and low-lying valley in each aeration period in pH profiles[76].

A typical profile for temperature descent rapidly as influent was mixed with the treated wastewater remaining in the reactor from the previous cycle. The temperature hit bottom soon after the fill phase. Following this bottom, temperature increased due to microbial activity. At approximately 250 min, the rate of increase in temperature has no more obvious change and continued thus for the rest of the cycle. Figure 6 shows the temperature change in a whole cycle.

As found, during the early stage of the cycle, temperature increased harder than the later, it is probably because the pollutants concentration differs: higher level pollutants concentration results in more activity of microorganisms which is the main reason for temperature change. The overall variation of sewage temperature in a certain cycle is relatively small, which is greatly influenced by the heat conduction and microbial metabolism of the

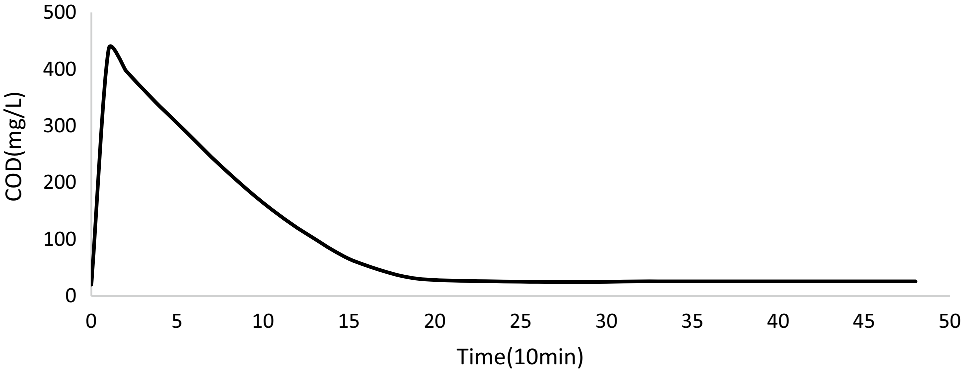

Figure 4. Typical profile for COD.

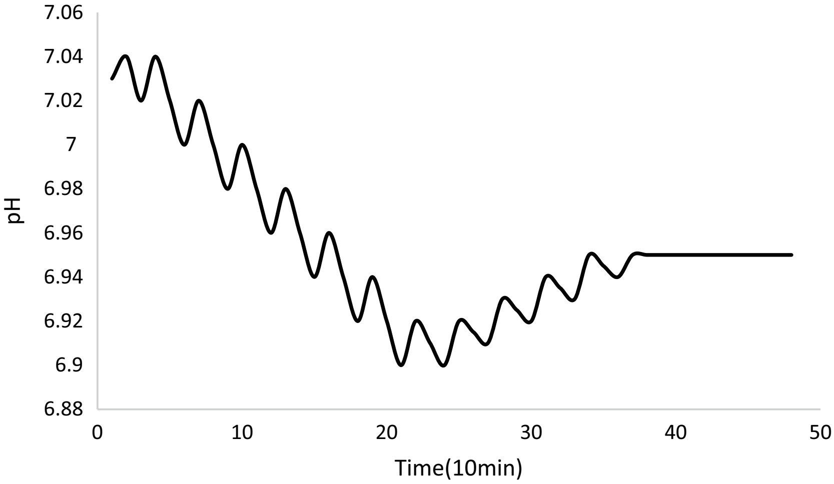

Figure 5. Typical profile for pH.

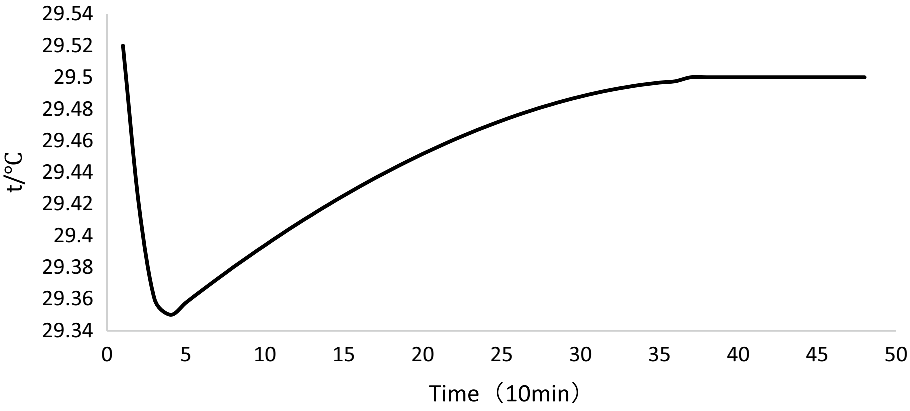

Figure 6. Typical profile for temperature.

environment, while the mechanical heat transfer, mainly by the pumps and aerators, has little influence on the variation of sewage treatment temperature[77].

Application. According to the principle of RFR model, the modeling process is divided into four steps as follows: (1) the collection of sample data; (2) the determination and ranking of the importance of features; (3) different number of features were added to the random forest model in order to select proper quantity of important features; (4) RFR model applied in practice.

Based on the operation process data, RFR soft sensor model is used to establish the COD prediction model of SBR effluent, which realizes the rapid prediction of effluent quality and provides the basis for the efficient and stable operation of the wastewater treatment process as shown in Fig. 7.

Assessed Input Variables. In this case, the temperature values was observed to increase with COD reduction and was considered useful in identifying the end of COD removal.

In the early stage of biochemical reaction, the anabolism of microorganism is intense, which produces an amount of CO_2. The quantity of CO_2 caused by anabolism is obviously more than that by aeration according to the result of measurement (Fig. 5) in the early stage. Moreover, the organic matter produces organic acid, which makes the pH value decrease further. Less residual organic matter caused lower production of CO_2 and organic acids, and the predominance of denitrification in the medium and end stage during this time period contribute to an overall increase in pH value. So the pH values were observed to decrease or increase according to different organic matter and was considered useful in identifying the residual quantity of COD.

A number of unprocessed and processed input variables such as pH, temperature, pH and temperature change in adjacent measurements, etc. were constructed and added to the set of independent variables. The selected processed input variables were constructed using the profile features.

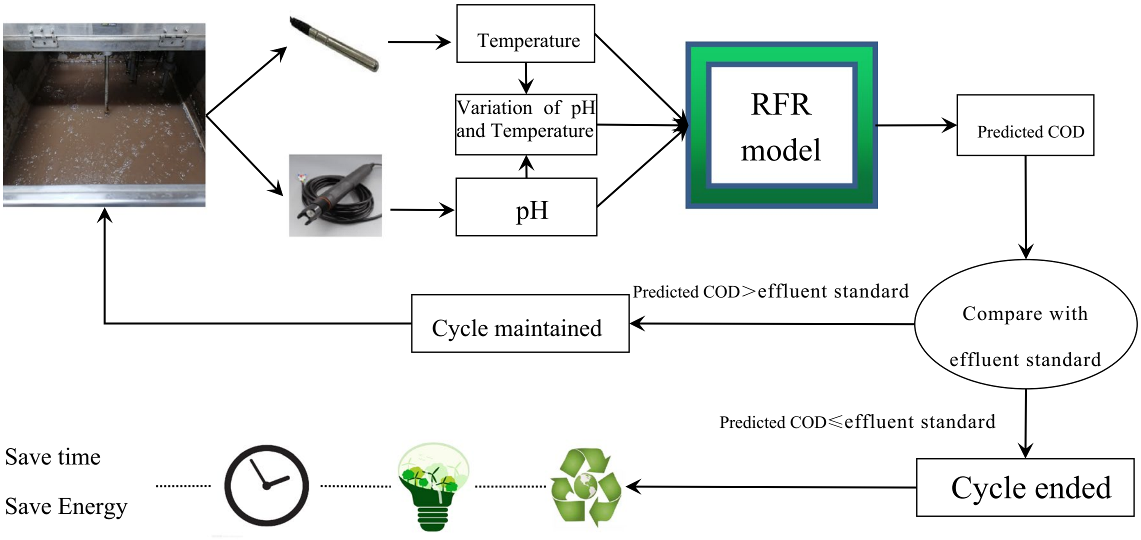
Temperature
Variation of pH and Temperature
pH
RFR model
Predicted COD
Predicted COD>effluent standard
Compare with effluent standard
Cycle maintained
Predicted COD≤effluent standard
Cycle ended
Save time
Save Energy

Figure 7. The technical balance between the SBR procedures and RFR algorithm.

The advantage of that pH and temperature were taken as unprocessed variables is it's simple and easy to detect them. Besides sensors for pH and temperature are not only low-cost but also have satisfying measurement accuracy.

Each variable set included a unique collection of input variables (Table 4). Within each 480 min cycle, data collected 0 ~ 30 min and 361 ~ 480 min were excluded to eliminate the effects of filling and settlement periods (as these phases were not part of the biological reaction phases of the treatment cycle).

Data from 40 treatment cycles were collected, 12 (30%) of which were randomly separated for use as a test dataset, and the remainder were used as a training dataset.

Assessment of RFR soft sensor model. The effectiveness of the RFR soft sensor model was assessed across 5 criteria. The effluent standard value of COD was set at 30 mg/l. Effluent standard value can vary due to local regulations. The assessment criteria are listed in Table 5.

Input variables	Description
pH	Raw pH data
ΔpH	$\Delta pH = pH_i - pH_{i-1}$, the difference between current pH value and previous
pH_{av}	Moving average of pH over the previous 10 records
$pH_{apex\text{-}nadir}$	pH apex value minus pH nadir value for each aeration period
T	Raw temperature data
ΔT	$\Delta T = T_i – T_{i-1}$, the difference between current t value and previous
T_{av}	Moving average of T over the previous 10 records
pH·T	pH multiply by T
$pH_{av} \cdot T_{av}$	pH_{av} multiply by T_{av}
T/pH	T divided by pH
T_{av}/pH_{av}	T_{av} divided by pH_{av}
ΔT·ΔpH	ΔT multiply by ΔpH

Table 4. Input variables.

Criterion	Description	Practical application
R^2	Referred to as the coefficient of determination, it is an indicator of the strength of the relationship between variables	Measures the strength of the relationship between predicted COD trend and actual trend
MSE	Mean square error (MSE) is a standard statistical metric to measure model performance; it measures the difference between sample and predict values and is a good measure of accuracy. The lower the MSE value the more accurate the prediction	Measures the average accuracy of the predicted COD trend against the actual trend
Percentage of COD removal	This criterion returns the percentage COD removal from the peak true concentration of the measured treatment cycle (30 ~ 360 min) to the predicted COD concentration which is below the effluent standard value for the first time	Provides a comparison of the COD concentration at which the cycle would have been ended by the model during a controlled cycle and the COD peak concentration at the beginning of a cycle
Percentage of time saved (T_{save})	$T_{save} = (330 - T_{thres})/330$ where T_{save} is the time saving (%), T_{thres} is the time at which the cycle would be ended by the model in a controlled scenario and 330 is the fixed time cycle length (min) set in an uncontrolled scenario	Indicates the time saved with the selected cut-off threshold value In general, the greater the time saved, the more the energy saved only if the accuracy is met the requirement
Accuracy	When the predicted COD concentration is below the effluent standard value for the first time, if it is true (predicted COD > measured COD) the accuracy meets the requirement. Symbol " + " for meeting the requirements, otherwise " – "	Indicates the accuracy at the cut-off threshold value

Table 5. Criteria of assessment.

Parameters	Average influent (mg/L)	Average effluent (mg/L)	Average removal (%)
COD	305	23	92.46
TP	0.97	0.10	89.69
NH_3-N	20.6	1.1	94.66

Table 6. Average influent and effluent results.

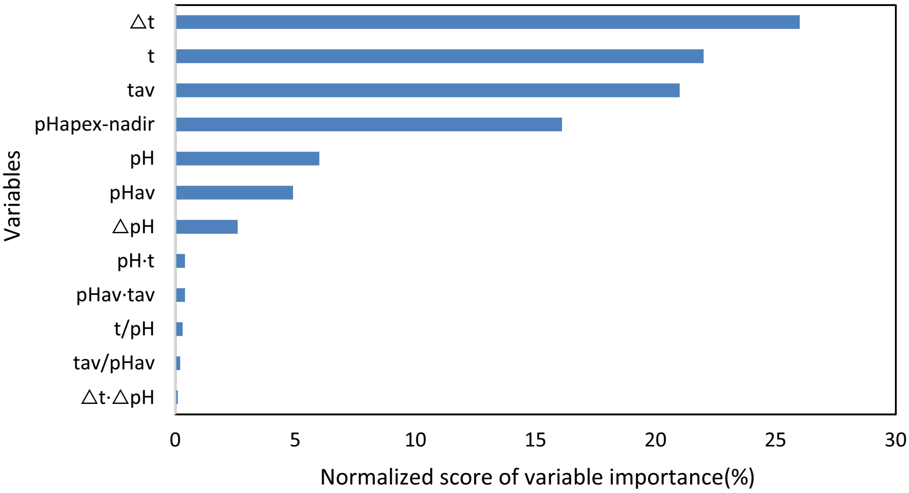

Figure 8. Ranking of each variable importance score.

Results

Average influent and effluent results.

Table 6 shows the related parameters of influent and effluent.

Ranking of variables.

Pearson correlation coefficient is a statistical measure used to determine the strength and direction of the linear relationship between two variables. Essentials of application of Pearson correlation coefficient in variables correlation ranking are: (1) Pearson correlation coefficient is commonly used in multiple regression analysis to select the most significant independent variables by calculating the correlation coefficients between each independent variable; (2) The correlation coefficient ranges from − 1 to 1, and the larger the absolute value, the stronger the correlation; (3) When the correlation coefficient value is close to 0, it indicates that the correlation between the two variables is very weak and they can be considered independent.

In order to guarantee the training effect of the RFR, the Pearson coefficient method was used to study the correlation of the subjects, and the variables with weak correlation were deleted. In order to prevent the occurrence of invalid variables, avoid overfitting and improve the training performance of the model, any variable with a normalized Pearson correlation coefficient value, that is regarded as the normalized score of variable importance, less than 0.01 was removed. The resulting normalized score of variable importance ordering diagram shows the 12 factors affecting COD concentration (Fig. 8). It was found that ΔT had the greatest influence on COD concentration, followed by T, T_{av}, $pH_{apex-nadir}$ and etc.

Data from 40 treatment cycles were collected. 70% of the whole data (28 treatment cycles) are randomly selected as training set for RFR model and 30% (12 treatment cycles) are selected as testing set to verify the accuracy of the model.

Based on the ranking of variable importance scores, it is evident that temperature-related variables hold the top three positions. Therefore, it can be concluded that temperature-related variables play a dominant role in the data analysis. Some studies have shown that the metabolic activity of microbial communities in wastewater treatment bioreactors can cause an increase in water temperature[78,79]. This is because the microorganisms in the reactor produce a large amount of heat through the degradation and metabolism of organic matter, leading to an increase in the temperature inside the reactor. Furthermore, it should be noted that while pH is indeed a contributing factor, its significance is not as strong as that of pHapex-nadir. pHapex-nadir, which is calculated by pH apex value minus pH nadir value for each aeration period, effectively quantifies the amount of carbon dioxide generated by microbial activity during a 20 min agitation.

Variables definition.

In order to select the variables set, different numbers of variables were selected according to the importance of variables, and then were added to the RFR model, as shown in Fig. 9. It was found that when the top 7 variables were selected, the R^2 of training set and the test set did not increase and the MSE did not decrease obviously, so the top 7 variables were selected as the variable of the optimized RF model, specific as follows: ΔT, T, Tav, pHapex-nadir, pH, pHav and ΔpH.

Predict results by RFR.

Figure 10 shows the comparison of predicted and measured COD concentrations on the test set.

The COD degradation trend, as well as the deviation between predicted and measured values, can be observed from the variations in the curves depicted in Fig. 10. The predicted values enable a rough estimation of the processing effect and level of pollutant degradation within a single cycle. Although the accuracy between true and predicted values may not be perfect, the slight discrepancy only exists in the initial stage of the process and soon disappears.

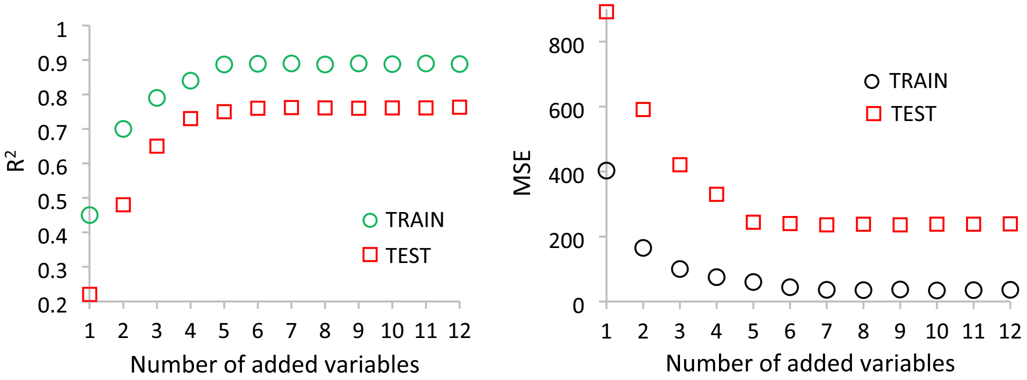

Figure 9. Evaluation indexes with different quantitative of variables.

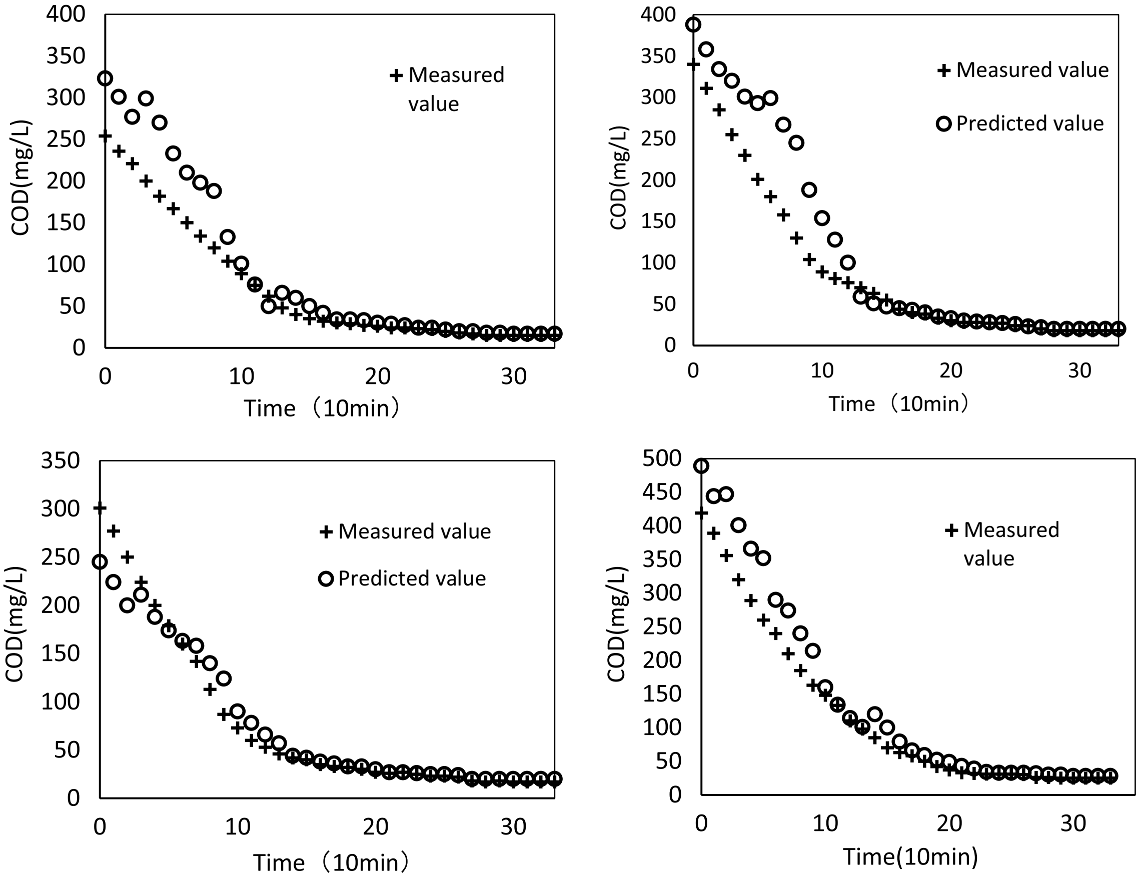

Figure 10. Comparison of predicted and actual/measured COD concentrations.

In the process of sewage treatment, the change of COD is influenced by various uncertain factors in operating conditions. These factors can cause significant differences in the accuracy of prediction of COD during different stages of processing. In the early stage, these uncertain factors have a stronger influence, which results in a obvious error between the predicted and measured values; with the passage of time, the processing conditions tend to stabilize, and the impact of uncertain factors on COD changes gradually decreases, leading to a reduction in the error between predicted and measured values.

Therefore, in the proposed methodology, the magnitude of the error between predicted and measured values is mainly affected by the processing stage. In the early stage, the error may be relatively large, but as time progresses, the error will gradually decrease and eventually reach a more accurate prediction effect.

The RFR soft sensor model output, serving as the predicted value of water quality in the given scenario, can be instrumental in optimizing the wastewater treatment process. This can be achieved by reducing energy consumption and enhancing the efficiency of chemical and biological processes. Specifically, if the predicted

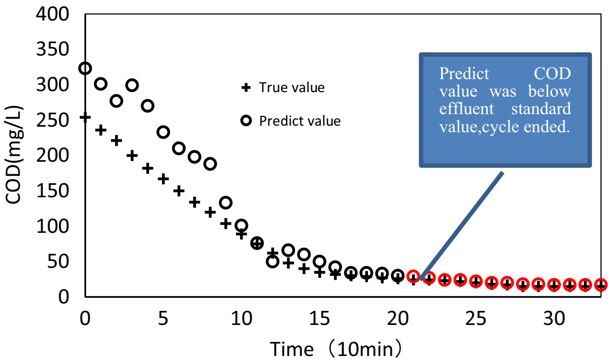

Figure 11. Condition of cycle ends.

Criterion	Test set												Average
	1	2	3	4	5	6	7	8	9	10	11	12	
R^2	0.77	0.7	0.72	0.79	0.76	0.75	0.84	0.77	0.77	0.71	0.78	0.79	0.763
MSE	264	268	245	247	218	197	199	235	270	265	240	216	239
Percentage of COD removal (%)	88.6	91.5	91	93.4	90	87.7	88.1	93.8	92.1	90.4	92.7	93.6	91.075
Percentage of time saved (%)	36.4	33.3	36.4	9.1	36.4	39.4	33.3	6.1	12.1	18.2	18.2	12.1	24.25
Accuracy	+	+	+	+	+	−	+	+	+	+	+	+	Percentage of hits 91%

Table 7. Results of assessment.

COD value falls below the effluent standard level, the process can transition into settlement mode immediately, with the agitator and aerator being switched off, thus bringing the cycle to a close. This method allows for cycle completion to be controlled by an artificial intelligence and automatic control system, as opposed to a fixed-time control approach that lacks precision. This is illustrated in Fig. 11.

Table 7 shows the assessment results of the test set 1–12.

Discussion

Benefits of the RFR soft sensor model. RFR is a machine learning model used for predictive analytics tasks, particularly for regression problems. RFR is an ensemble learning method that combines multiple decision tree models to create a more robust and accurate predictor.

The RFR algorithm randomly selects subsets of the input variables and samples of the training data to construct decision trees, which are then combined into a forest. During prediction, the RFR model aggregates the output of individual decision trees to produce a final prediction. This approach helps to reduce the impact of overfitting and improves the model's performance on output data.

In the context of wastewater treatment plants, RFR *soft sensor* model can be used to predict water quality (COD) through simpler diameters, in this way complex and expensive sensors will be replaced. Although a genuine COD sensor may be an option, several reasons or factors can result in its unsuitability and cause certain issues to arise: (1) Genuine COD sensors usually cost much; (2) Due to the presence of suspended solids in sewage, the COD value measured by genuine COD sensors can be unstable and exhibit significant fluctuations; (3) Some genuine COD sensors can detect organic compounds with a double bond sensitively while other organic compounds without a double bond are failed to be detected, so the error can not be ignored.

Moreover, the RFR model overcomes the shortcomings of slow convergence speed and large number of samples requiring of neural networks. Neural networks are powerful models that can learn complex patterns in data. However, training a neural network can be computationally expensive and require a large amount of data. In particular, deep neural networks or large-scale networks may take a long time to converge during training due to the sheer number of parameters that need to be learned through multiple iterations.

In contrast, RFR model are composed of multiple decision tree models, each trained on a random subset of the data. This approach has advantages: (1) RFR model does not require as much data as neural networks, since each decision tree model can work well with smaller datasets; (2) RFR model can be easily parallelized, which means they can be trained more quickly than neural networks on multi-core computer systems.

Comparison RFR soft sensor model with others. Comparing with methods used in pollutant removal technology (Table 1), the proposed methodology requires only two types of raw data that are easy to obtain, greatly reducing the workload of data acquisition. The weakness of the proposed methodology is that prediction value is not so accurate between measured and predicted value at the first stage of the progress. Additionally, even if the R^2 and MSE of RFR model are not satisfactory, it performs well in predicting accuracy at the cut-off threshold value, as shown in Table 7, and this is very concerned in the field of engineering.

Potential trade-offs or unintended consequences. In the practical application of the proposed methodology, it is possible that the COD value meet the standard while other indicators such as ammonia or phosphorus do not. To address this issue, relationship models can be established using pH and temperature as variables to predict the other parameters. However, this approach is limited to artificial intelligence methods only. In addition, an empirical judgment system can be established, such as the sewage treatment time generally being within a certain range, if predicted results exceeds this range, the output results of the proposed methodology are deemed to require modification.

Methods or options for improvement. Increase conductivity or other readily available parameters as input variables to improve prediction accuracy. The following are detailed discussions:

Variables such as conductivity, MLSS, DO and ammonia can also be used as premises to predict COD. The impact to the accuracy and efficiency of the proposed methodology may be: (1) In general, in the sewage treatment process, the electrical conductivity of the solution shows a trend of decreasing gradually, which is related to the decrease of COD value, hence the electrical conductivity may improve the accuracy and efficiency; (2) MLSS should show a trend of increasing gradually, however, the change of MLSS is not obvious in one cycle (480 min). Furthermore, the accuracy of MLSS sensor is easily affected by the color of wastewater, this will obviously increase the uncertainty of the data measured by the sensors; (3) The SBR works according to aeration-agitation periodicity, DO presents increase–decrease periodicity change, which obviously has no correlation with COD value change trend; (4) During the sewage treatment process, the ammonia concentration in the solution generally exhibits a gradual decrease, similar to the trend observed in COD. However, in some cases, such as a lack of dissolved oxygen that inhibits nitrification, there may be no significant reduction in the ammonia value even when COD is reduced. As a result of the non-synchronous nature of the changes in these two parameters, predicting COD using ammonia as a variable may introduce uncertainty into the analysis.

Encrypt the frequency of data acquisition, such as collecting data every 5 minutes, then it can be five minutes in advance to predict, which further improves the efficiency of the proposed methodology.

Add ammonia and phosphorus as prediction targets to balance organic and inorganic wastewater indicators and improve practicality.

Conclusions and outlook

Simple and stable sensors (pH, temperature) were utilized to predict COD values throughout the process. The RFR model employed in the study can be regarded as a "soft sensor", which assists in monitoring the treatment effect.

The SBR was optimized using artificial intelligence and an automatic control system to increase automation, as well as save both time and energy. pH and temperature sensors collected data, which were input into the RFR model, the model then outputted real-time COD values. Once the predicted COD value fell below the effluent standard value, the cycle ended by cutting down the agitator and aerator, and the process entered the settlement mode directly. The proposed methodology replaced fixed-time control, which was uncontrolled. In 12 test cases, the percentage of COD removal (%) was about 91. 075, while an average of 24. 25% of time or energy was saved. These results demonstrate that this approach can increase treatment capacity and reduce energy consumption, representing a low-carbon technology.

R^2 on the test set is around 0.791, although it is not too high, but the accuracy at the cut-off threshold value of COD is around 91% which is acceptable for the prediction. It is quite simple and almost accurate to acquire the processing effect and the level of degradation of pollutants at anytime. Although it is not so accurate between true and predict value, but the embarrassment only occured at the first half of the progress and it soon vanished. The accuracy of the medium and end stage is more important than that of the early stage, the reason for the above fact is explained below. Artificial intelligence and automatic control system leaded to a optimized way but the satisfied accuracy of predict COD value is prerequisite. Basing on the fact, accuracy requirements are different at each stage in a controlled scenario: in the medium and end stage, especially when approaching the stage of effluent standard compliance, greater emphasis is placed on precision and accuracy. However, in the early stage, the accuracy does not significantly affect the control strategy.

Due to the non-linearity and uncertainty of the variation of pH value with time in SBR process, predict results are unstable because of different algorithm and over-fitting by ANN method. Due to the parallel information distribution and storage of structural preprocessing, RFR has strong fault tolerance and the ability to adapt to the external environment through learning. The ability of pattern recognition and comprehensive reasoning undoubtedly opens up a broad prospect for experimental research.

One limitation of this research is its exclusive focus on SBR methodology. However, there exists the potential to modify the procedure to cater to other technologies, particularly batch wastewater treatment systems. By increasing the frequency of data acquisition, such as collecting data every 5 minutes, it may be possible to predict factors up to five minutes ahead of time, thereby further enhancing the efficiency of the proposed methodology. To improve its practicality, ammonia and phosphorus could be included as prediction targets, as this would help balance organic and inorganic wastewater indicators.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 22 February 2023; Accepted: 1 June 2023
Published online: 05 June 2023

References

1. Chen, P., Zhao, W., Chen, D., Huang, Z. & Zhang, C. Research progress on integrated treatment technologies of rural domestic sewage: A review. *Water* **14**(15), 2439–2439 (2022).
2. Wang, C. *et al.* Revealing factors influencing spatial variation in the quantity and quality of rural domestic sewage discharge across China. *Process Saf. Environ. Prot.* **162**, 200–210 (2022).
3. Srivastava, G. *et al.* Influence of variations in wastewater on simultaneous nutrient removal in pre-anoxic selector attached full-scale sequencing batch reactor. *Int. J. Environ. Sci. Technol.* **1**, 1–18 (2022).
4. Masłoń, A. Impact of uneven flow wastewater distribution on the technological efficiency of a sequencing batch reactor. *Sustainability* **14**(4), 2405 (2022).
5. Alhazmi, H. E., Yin, Z., Grubba, D., Majtacz, J. B. & Mąkinia, J. Comparison of the efficiency of deammonification under different DO concentrations in a laboratory-scale sequencing batch reactor. *Water* **14**(3), 368 (2022).
6. Chen, D. & Li, H. Enhanced simultaneous partial nitrification and denitrification performance of aerobic granular sludge via tapered aeration in sequencing batch reactor for treating low strength and low COD/TN ratio municipal wastewater. *Environ. Res.* **209**, 112743 (2022).
7. Schwitalla, P. *et al.* NH_4^+ ad/desorption in sequencing batch reactors: simulation, laboratory and full-scale studies. *Water Sci. Technol.* **58**(2), 345 (2008).
8. Kauder, J., Boes, N., Pasel, C. & Herbell, J. D. Combining models ADM1 and ASM2d in a sequencing batch reactor simulation. *Chem. Eng. Technol.* **30**(8), 1100–1112 (2007).
9. Wang, J., Zhao, X., Guo, Z., Yan, P. & Gao, X. A full-view management method based on artificial neural networks for energy and material-savings in wastewater treatment plants. *Environ. Res.* **211**, 113054 (2022).
10. Alharbi, M., Hong, P.-Y., Laleg, K. & Taous, M. Sliding window neural network based sensing of bacteria in wastewater treatment plants. *J. Process Control* **110**, 35–44 (2022).
11. Salim, H., Hilal, L. & Samir, F. Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: A comparative study. *Environ. Process.* **3**(1), 153–165 (2016).
12. Sharghi, E., Nourani, V., Ashrafi, A. A. & Gökçekuş, H. Monitoring effluent quality of wastewater treatment plant by clustering based artificial neural network method. *Desalin. Water Treat.* **164**, 86–97 (2019).
13. Hasanlou, H., Abdolabadi, H. & Aghashahi, M. Application of factor analysis in a large-scale industrial wastewater treatment plant simulation using principal component analysis-artificial neural network hybrid approach. *Environ. Prog. Sustain. Energy* **34**(5), 1322–1331 (2015).
14. Sadek, A. H., Fahmy, O. M., Mahmoud, N. & Mostafa, M. K. Predicting Cu (II) adsorption from aqueous solutions onto nano zero-valent aluminum (nZVAl) by machine learning and artificial intelligence techniques. *Sustainability* **15**(3), 2081 (2023).
15. Faiz, A. A., Mahmoud, N., Ismail, R. & Faizal, B. Artificial neural network and techno-economic estimation with algae-based tertiary wastewater treatment. *J. Water Process Eng.* **40**, 101761 (2021).
16. Mohamed, G. A. & Mahmoud, N. Treatment of water contaminated with diazinon by electro-Fenton process: Effect of operating parameters, and artificial neural network modeling. *Desalin. Water Treat.* **182**, 277–287 (2020).
17. Mithil, K. N. *et al.* Artificial neural network and cost estimation for Cr(VI) removal using polycationic composite adsorbent. *Water Environ. J.* **34**(S1), 29–40 (2019).
18. Fmahmoud, N., Karam, M., Michael, A. & Ibrahim, M. G. Chapter 9-sustainable management of wastewater treatment plants using artificial intelligence techniques. *Soft Comput Tech. Solid Waste Wastewater Manag.* **1**, 171–185 (2021).
19. Nashia, D. *et al.* Artificial intelligence and multivariate statistics for comprehensive assessment of filamentous bacteria in wastewater treatment plants experiencing sludge bulking. *Environ. Technol. Innov.* **19**, 1 (2020).
20. Alberto, P. *et al.* Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing* **214**, 242–268 (2016).
21. Emon, M. & Mohaiminul, I. An overview of artificial neural network. *Am. J. Comput. Sci. Appl.* **3**(16), 1 (2019).
22. Kudus, S. A. *et al.* An overview current application of artificial neural network in concrete. *Adv. Mater. Res.* **626**(626), 372 (2012).
23. Zelin, L. *et al.* Application of artificial neural networks in global climate change and ecological research: An overview. *Chin. Sci. Bull.* **55**(34), 3853–3863 (2010).
24. Bhattacharjee, A., Murugan, R., Soni, B. & Goel, T. Ada-GridRF: A fast and automated adaptive boost based grid search optimized random forest ensemble model for lung cancer detection. *Phys. Eng. Sci. Med.* **45**(3), 981–994 (2022).
25. Sun, D., Xu, J., Wen, H. & Wang, Y. An optimized random forest model and its generalization ability in landslide susceptibility mapping: Application in two areas of Three Gorges Reservoir China. *J. Earth Sci.* **31**(6), 1068–1086 (2020).
26. Mateo, P. V., Mesa, F. J. M., Villanueva, B. J. & Alonso, Á. C. A Random forest model for the prediction of FOG content in inlet wastewater from urban WWTPs. *Water* **13**(9), 1237 (2021).
27. Pengxiao, Z. *et al.* A random forest model for inflow prediction at wastewater treatment plants. *Stoch. Env. Res. Risk Assess.* **33**(10), 1781–1792 (2019).
28. Olshen, R. A., Leo, B. & Jerome, F. Classification and regression trees. *Chapman and Hall Press:London, UK* **1**, 10 (2017).
29. Ferrante, M., Demarco, P. & Origgi, D. OD177-Random forest regression on CT data to predict effective dose and class of effective dose in compliance to the new Italian regulation. *Phys. Med.* **92**, S138 (2021).
30. Srimathi, S., Yamuna, G. & Nanmaran, R. Threshold based stochastic regression model with gabor filter for segmentation and random forest classification of lung cancer. *J. Comput. Theor. Nanosci.* **16**(4), 1666–1673 (2019).
31. Cristina, Z. Modeling the connection between bank systemic risk and balance-sheet liquidity proxies through random forest regressions. *Admin. Sci.* **10**(3), 52 (2020).
32. Yihui, C. & Minjie, L. Evaluation of influencing factors on tea production based on random forest regression and mean impact value. *Agric. Econ.* **65**, 340–347 (2019).
33. Luo, Y., Yan, J., Mcclure, S. C. & Li, F. Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. *Environ. Sci. Pollut. Res. Int.* **29**(22), 33205–33217 (2022).
34. Song, H. E., Jianhua, W. U., Wang, D. & Xiaodong, H. E. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere* **290**, 133388 (2021).
35. Zhangwen, S. U., Lin, L., Chen, Y. & Hu, H. Understanding the distribution and drivers of PM2.5 concentrations in the Yangtze River Delta from 2015 to 2020 using random forest regression. *Environ. Monit. Assess.* **194**(4), 284 (2022).
36. Wang, Q. *et al.* Spatially explicit reconstruction of the population distribution in the Tuojiang River Basin during 1911–2010 using random forest regression. *Reg. Environ. Change* **22**(1), 1 (2022).
37. Wang, F. *et al.* Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environ. Res.* **202**, 111660 (2021).
38. Shijun, C. *et al.* Medium and long-term runoff forecasting based on a random forest regression model. *Water Supply* **20**(8), 3658–3664 (2020).

39. Harrison, J. W., Lucius, M. A., Farrell, J. L., Eichler, L. W. & Relyea, R. A. Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression. *Sci. Total Environ.* **763**, 143005 (2020).
40. Weiran, Y. *et al.* Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environ. Pollut.* **245**, 746–753 (2018).
41. Kronholm, S. C., Capel, P. D. & Terziotti, S. statistically extracted fundamental watershed variables for estimating the loads of total nitrogen in small streams. *Environ. Model. Assess.* **21**(6), 681–690 (2016).
42. Shi, G. Y. *et al.* Modeling the response of negative air ions to environmental factors using multiple linear regression and random forest. *Eco. Inform.* **66**, 1 (2021).
43. Keramat, J. M., Saeid, M. S., Mousazadeh, H., Ghasemi, V. M. & Rahimi, M. M. Real-time moisture ratio study of drying date fruit chips based on on-line image attributes using kNN and random forest regression methods. *Measurement* **10**, 8899 (2020).
44. Balogun, A. L. & Tella, A. Modelling and investigating the impacts of climatic variables on ozone concentration in Malaysia using correlation analysis with random forest, decision tree regression, linear regression, and support vector regression. *Chemosphere* **299**, 134250 (2022).
45. Wang, J. *et al.* A full-view management method based on artificial neural networks for energy and material-savings in wastewater treatment plants. *Environ. Res.* **211**, 113054 (2022).
46. Fernandez de Canete, J. & Del Saz-Orozco, P. Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network. *Expert Syst. Appl.* **63**, 8–19 (2016).
47. He, F., Wang, J. & Chen, W. Cleaner production assessment for wastewater treatment plants based on backpropagation artificial neural network. *Neuro Quantol.* **16**(6), 1 (2018).
48. Zhao, H. Y., Huang, F. L., Li, L. & Zhang, C. Y. Optimization of wastewater anaerobic digestion treatment based on GA-BP neural network. *Desalin. Water Treat.* **122**, 30–35 (2018).
49. Junfei, Q., Gaitang, H., Honggui, H. & Wei, C. Wastewater treatment control method based on a rule adaptive recurrent fuzzy neural network. *Int. J. Intell. Comput. Cybern.* **10**(2), 94–110 (2017).
50. Watanabe, S. & Yamana, H. Overfitting measurement of convolutional neural networks using trained network weights. *Int. J. Data Sci. Anal.* **14**(3), 261–278 (2022).
51. Kai, I., Yuta, O., & Moriya, N. Overfitting of artificial-neural-network-based nonlinear equalizer for multilevel signals in optical communication systems. OPTO conference, **2020**.
52. Oyedotun, O. K., Olaniyi, E. O. & Khashman, A. A simple and practical review of over-fitting in neural network learning. *Int. J. Appl. Patt. Recognit.* **4**(4), 307–328 (2017).
53. Buskirk, T. D. Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Surv. Pract.* **11**(1), 1–13 (2018).
54. Matthias, S. & Rosie, Y. Z. The random forest algorithm for statistical learning. *Stand. Genomic Sci.* **20**(1), 3–29 (2020).
55. Kim, B., Cha, J. W., Chang, K. & Lee, C. Visibility prediction over South Korea based on random forest. *Atmosphere* **12**(5), 552 (2021).
56. Massimo, S. *et al.* A random forest approach to estimate daily particulate matter, nitrogen dioxide, and ozone at fine spatial resolution in Sweden. *Atmosphere* **11**(3), 1 (2020).
57. Rubal, D. K. Evolving Differential evolution method with random forest for prediction of Air Pollution. *Proc. Comput. Sci.* **132**, 824–833 (2018).
58. Shamsoddini, A., Aboodi, M. R. & Karami, J. Tehran air pollutants prediction based on random forest feature selection methodisprs. *Int. Arch. Photogram. Remote Sens. Spatia* **4**, 483–488 (2017).
59. Shi, G.-Y. *et al.* Modeling the response of negative air ions to environmental factors using multiple linear regression and random forest. *Ecol. Inf.* **66**, 1 (2021).
60. Min, J. S. *et al.* Identification of primary effecters of N2O emissions from full-scale biological nitrogen removal systems using random forest approach. *Water Res.* **184**, 116–144 (2020).
61. Vitorino, D. *et al.* A random forest algorithm applied to condition-based wastewater deterioration modeling and forecasting. *Proc. Eng.* **89**, 401–410 (2014).
62. Buras, M. P. & Solano, D. F. Identifying and estimating the location of sources of industrial pollution in the sewage network. *Sensors* **21**(10), 3426 (2021).
63. Medl, M., Rajamanickam, V., Striedner, G. & Newton, J. Development and validation of an artificial neural-network-based optical density soft sensor for a high-throughput fermentation system. *J. Processes* **11**(1), 297–307 (2023).
64. Cong, Q. M., Bo, G.-H. & Shi, H.-Y. Integrated soft sensor of COD for WWTP based on ASP model and RBF neural network. *J. Meas. Control* **56**(1–2), 295–303 (2023).
65. Hema, P. *et al.* Robust soft sensor systems for industry: Evaluated through real-time case study. *J. Meas. Sens.* **24**, 1 (2022).
66. Severino, A. G. V., De, L. J. M. M. & De, A. F. M. U. Industrial soft sensor optimized by improved PSO: A deep representation-learning approach. *J. Sens.* **22**(18), 6887–6899 (2022).
67. Dimitriadis, S. I. & Liparas, D. How random is the random forest?Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: From Alzheimer's disease neuroimaging initiative (ADNI) database. *Neural Regen. Res.* **13**(6), 962–970 (2018).
68. Jaime, L. S. *et al.* A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **134**, 93–101 (2019).
69. Christopher, S., Margaret, G. S., Chuck, E. B. & Anders, K. Semi-automated classification of exposed bedrock cover in British Columbia's Southern Mountains using a Random Forest approach. *Geomorphology* **285**, 214–224 (2017).
70. Hristos, T. & Georgia, P. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **11**(5), 910 (2019).
71. Kamakshaiah, M. & Prasada Rao, S. S. Applicability of random forests forecasting to international currency trade: An investigation through language. *Int. J. Bus. Anal. Intell.* **6**(1), 47–57 (2018).
72. Shu, F. Y., Joo, H. T. & Yu, L. Effect of substrate nitrogen/chemical oxygen demand ratio on the formation of aerobic granules. *J. Environ. Eng.* **131**(1), 86–92 (2005).
73. Tanwar, P., Nandy, T., Ukey, P. & Manekar, P. Correlating on-line monitoring parameters, pH, DO and ORP with nutrient removal in an intermittent cyclic process bioreactor system. *Biores. Technol.* **99**, 7630–7635 (2008).
74. Chang, C. H. & Hao, O. J. Sequencing batch reactor system for nutrient removal: ORP and pH profiles. *J. Chem. Technol. Biotechnol. Int. Res. Process Environ. Clean Technol.* **67**, 27–38 (1996).
75. Tong, Q., Mao, Z., & Sun, L. Variation of DO and pH value in the process of landfill leachate treatment by SBR. *Proc. Annu. Meet. China Silicate Soc. Environ. Protect.* 325–328 (2007).
76. Shane, F., Mcdermott, J., Doherty, E., Cooney, R. & Clifford, E. Application of neural networks and regression modelling to enable environmental regulatory compliance and energy optimisation in a sequencing batch reactor. *Sustainability* **14**, 4098 (2022).
77. Hao, X., Sun, S., Li, J. & Liu, R. Establishing and verification a temperature model for the process of water treatment. *J. Environ. Sci. (China)* **42**(12), 1–11 (2022).
78. Pochwała, S. & Kotas, P. Possibility of obtaining wastewater heat from a sewage treatment plant by the means of a heat pump: a case study. *E3S Web Conf.* **44**, 144–144 (2018).

79. Liang, J., Zhang, P., Cai, Y., Wang, Q. & Zhou, Z. Thermal effects. *Water Environ. Res. Publ. Water Environ. Fed.* **92**(10), 1406–1411 (2020).

Acknowledgements

The research is supported by Key Research and Development Project of Sichuan Provincial Science and Technology Plan, China (22ZDYF1876).

Author contributions

C.Q., C.Z. and Q.L. wrote the main manuscript text and all figures and tables. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.